Nucleic Acids Research, 1994, Vol. 22, No. 17 3495–3501

Database on the structure of large ribosomal subunit RNA

Peter De Rijk, Yves Van de Peer, Sabine Chapelle and Rupert De Wachter*
Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium

ABSTRACT

A database on large ribosomal subunit RNA is made available. It contains 258 sequences. It provides sequence, alignment and secondary structure information in computer-readable formats. Files can be obtained using ftp.

INTRODUCTION

This paper presents a comprehensive database of large ribosomal subunit RNA (further abbreviated as LSU rRNA) structures. Our goal is to offer researchers on-line acces to LSU rRNA sequences in the form of an alignment containing secondary structure information, in a format suitable for use in computer programs. Literature references and accession numbers in sequence databases are included as well as taxonomic information.

This database will conceivably be used to perform phylogenetic analysis, to find primers or probes, and to elicit the secondary structure of newly determined sequences. It can also be an invaluable tool to find sequence errors introduced during gel reading or typing. These errors cause anomalies in the sequence or structure which can often be easily detected by alignment to a set of known sequences and comparison of their possible secondary structure.

New entries or updates in the EMBL sequence database (1) are continuously scanned for new LSU rRNA sequences using the Current Sequence Awareness program (a service of the Belgian EMBnet Node). These sequences are used to update older entries, or added to the database as new entries. They are then aligned, and their secondary structure is investigated and incorporated into the alignment using the program DCSE (2). When anomalies or errors are found in the annotations in the sequence libraries, these are corrected when possible. A note indicating the changes made is added.

CONTENTS OF THE DATABASE

Only complete or reasonably complete sequences are being incorporated into the database. Partial sequences are excluded when the combined length of the sequence segments in *Escherichia coli* LSU rRNA homologous to the sequenced segments, amounts to less than 70% of the total *Escherichia coli* sequence. The database currently contains 258 sequences, viz. 42 eukaryal, 16 archaeal, 81 bacterial, 36 plastidial and 83 mitochondrial sequences.

Table 1 shows a list of species for which the LSU rRNA structure is recorded in the database. The same taxonomic classification is used as in the small ribosomal subunit rRNA database (3). For the domain Eukarya, the taxonomic classification of the species is according to Brusca and Brusca (4) for the Animalia, according to Cronquist (5) for the higher plants, according to Ainsworth *et al.* (6) for the zygomycetes and ascomycetes, according to Moore (7) for the basidiomycetes, and according to Margulis *et al.* (8) for the remaining eukaryotes, viz. the Protoctista.

For the Bacteria and the Archaea, the classification is based on the construction of evolutionary trees, explained into more detail in a previous compilation of small ribosomal subunit RNA sequences (9). In short, evolutionary trees are constructed by the neighbor-joining method (10) for all new sequences retrieved from the EMBL (1) and/or GenBank (11) nucleotide libraries. According to the phylogenetic position of the sequences, they are assigned to one of the taxa listed in Fig. 1 of a previous compilation (9) and described essentially by Woese and coworkers (12, 13) .

HETEROGENEITY IN SEQUENCE AND CHAIN LENGTH

Bacterial, archaeal and plastidial LSU sequences have a relatively constant length of approximately 2900 nucleotides. However, eukaryotic sequences show a much greater diversity, ranging in length from sizes comparable to those of the bacteria to over 5000 bases in the *Homo sapiens* sequence. The presence of extra nucleotides seems to be restricted mainly to several extremely variable areas, which occupy a constant position relative to the more conserved parts of the sequences (14,15). Sequence variation is even larger in mitochondria. The molecules found in animal and kinetoplastid mitochondria even miss large parts of the sequence conserved in other LSU rRNAs, and can be under 1000 nucleotides in size. Plant and fungi mitochondrial LSU rRNAs have chain lengths comparable to or larger than those found in bacteria.

SECONDARY STRUCTURE MODEL

Figures 1 and 2 show secondary structure models for a procaryotic (*Escherichia coli*) and a eukaryotic (*Saccharomyces cerevisiae*) LSU rRNA. A core structure is conserved in the majority of eukaryotic and bacterial LSU rRNAs. In the mitochondria of kinetoplastids and animals several helices of this core are absent. Other mitochodria have most of the helices of the core, although the structural variability is higher than among bacteria. The variable insertion regions in Eukaryotes can have

*To whom correspondence should be addressed

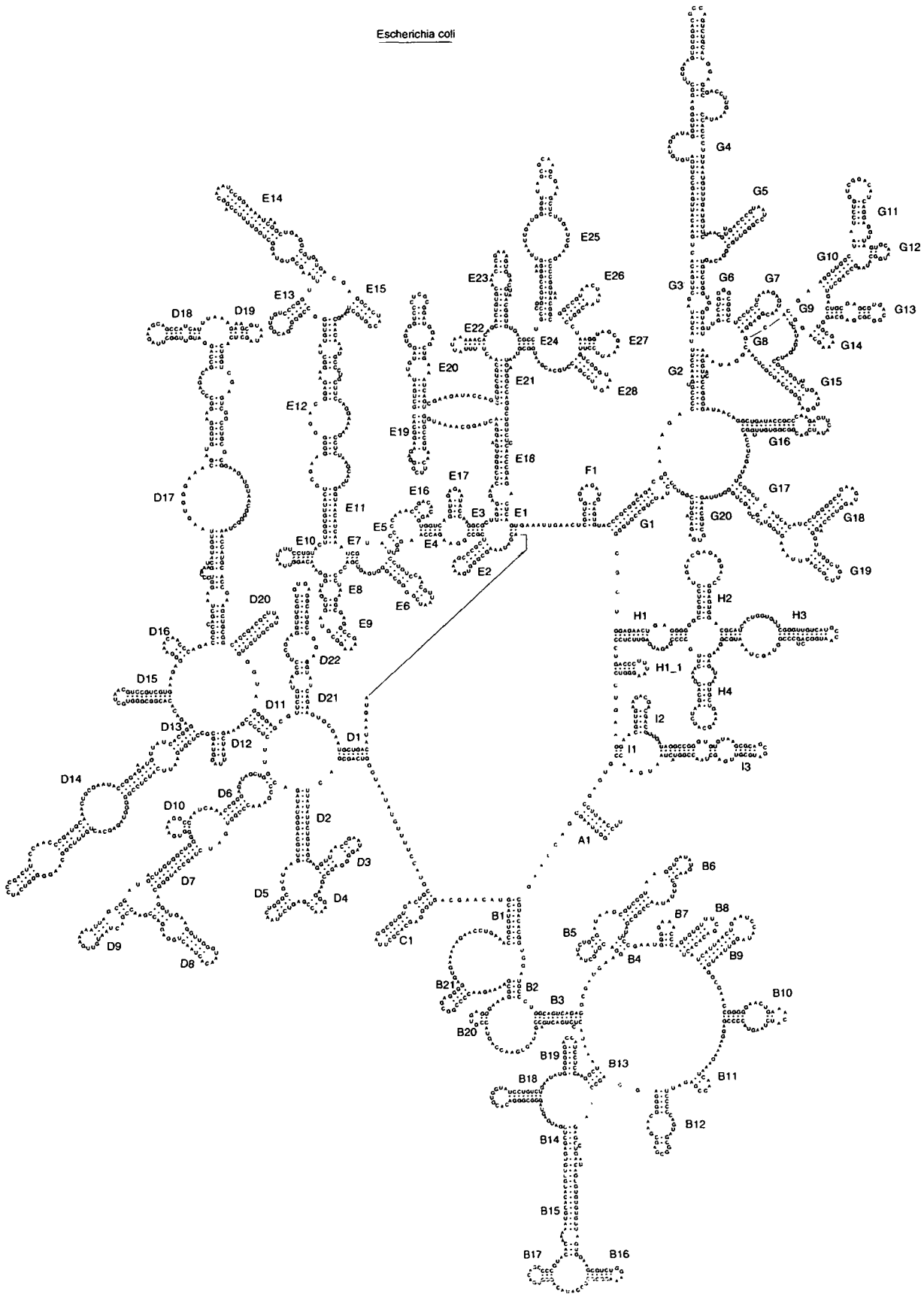

Figure 1. Secondary structure model for *Escherichia coli* LSU rRNA. The sequence is written clockwise from 5′ to 3′ terminus.

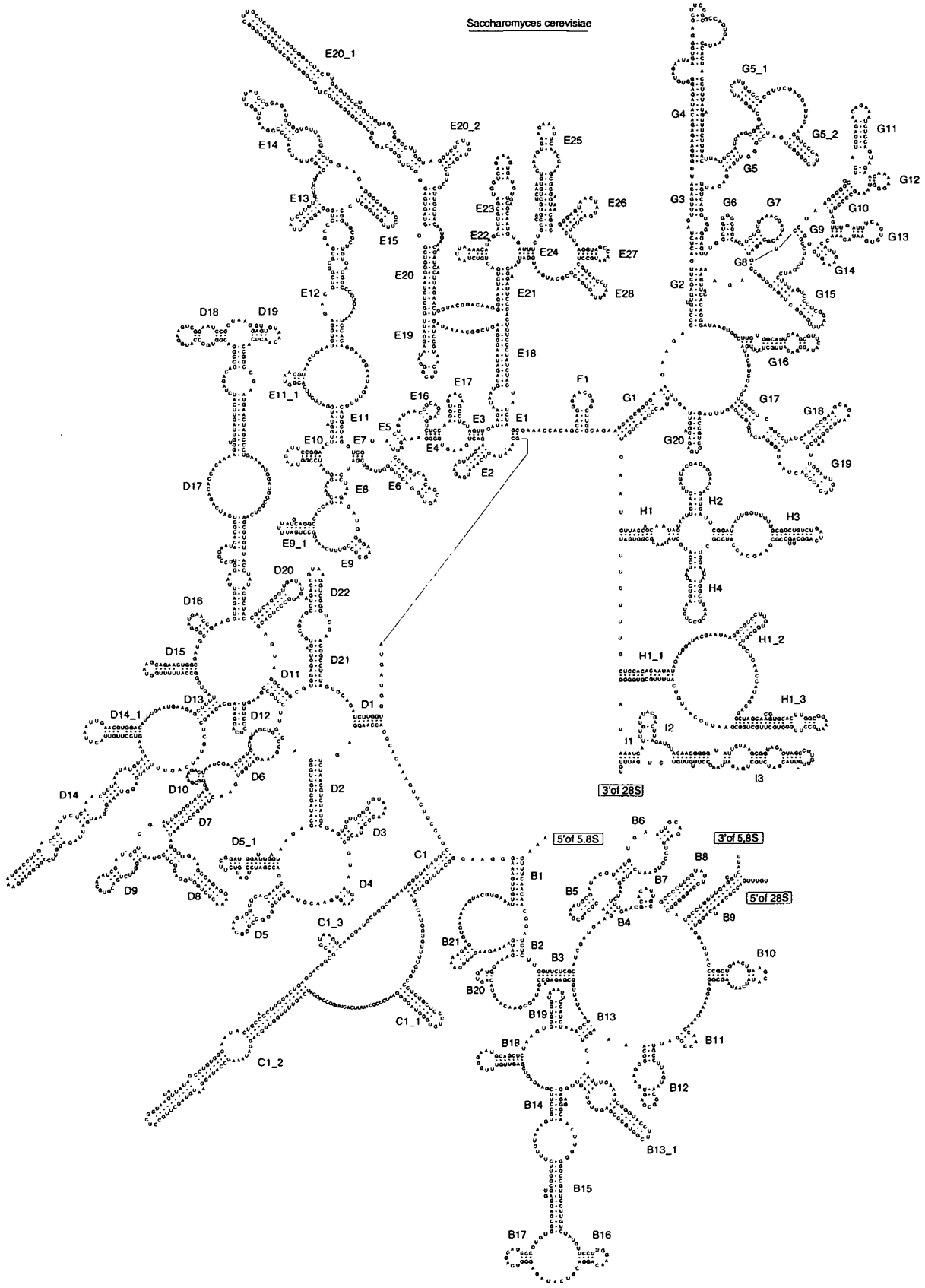

Figure 2. Secondary structure model for *Saccharomyces cerevisiae* LSU rRNA.

Table 1. List of species for which LSU rRNA structure is recorded in the database[a]

ARCHAEA

CRENARCHAEOTA
- Desulfurococcus mobilis
- Pyrobaculum islandicum
- Sulfolobus acidocaldarius
- Sulfolobus solfataricus
- Thermofilum pendens DSM 2475
- Thermoproteus tenax

EURYARCHAEOTA

HALOBACTERIA
- Halobacterium halobium
- Halobacterium maris-mortui
- Halococcus morrhuae ATCC 17082
- Haloferax volcanii

ARCHAEOGLOBALES
- Archaeoglobus fulgidus

METHANOBACTERIALES
- Methanobacterium thermoautotrophicum

METHANOCOCCALES
- Methanococcus vannielii

METHANOMICROBIUM GROUP
- Methanospirillum hungatei

THERMOCOCCALES
- Thermococcus celer

THERMOPLASMA
- Thermoplasma acidophilum

BACTERIA

PROTEOBACTERIA ALPHA
- Bradyrhizobium japonicum DSM 30131
- Rhodobacter capsulatus DSM 938
- Rhodobacter sphaeroides 1
- Rhodobacter sphaeroides 2
- Rhodobacter sphaeroides 3
- Rhodopseudomonas palustris DSM 126

PROTEOBACTERIA BETA
- Bordetella avium
- Bordetella bronchiseptica
- Bordetella parapertussis
- Bordetella pertussis
- Neisseria gonorrhoeae
- Neisseria meningitidis
- Pseudomonas cepacia DSM 50181

PROTEOBACTERIA GAMMA
- Aeromonas hydrophila
- Escherichia coli 1
- Escherichia coli 2
- Escherichia coli 3
- Escherichia coli 4
- Plesiomonas shigelloides NCIMB 9242
- Ruminobacter amylophilus
- Pseudomonas aeruginosa ATCC 10145
- Pseudomonas perfectomarina

PROTEOBACTERIA EPSILON
- Campylobacter jejuni ATCC 43431

GRAM POSITIVES AND RELATIVES, HIGH G+C
- Frankia sp. 1
- Frankia sp. 2
- Micrococcus luteus
- Mycobacterium kansasii ATCC 12478
- Mycobacterium leprae 1
- Mycobacterium leprae 2
- Streptomyces ambofaciens ATCC 23877
- Streptomyces griseus KCTC 9080

GRAM POSITIVES AND RELATIVES, LOW G+C
- Bacillus alcalophilus
- Bacillus anthracis
- Bacillus cereus NCTC 11143
- Bacillus globisporus DSM 4
- Bacillus lichenifomis DSM 13
- Bacillus sp.
- Bacillus stearothermophilus 1
- Bacillus stearothermophilus 2
- Bacillus subtilis 1
- Bacillus subtilis 2
- Clostridium botulinum 1
- Clostridium botulinum 2
- Clostridium botulinum 3 ATCC 25765
- Clostridium botulinum 4
- Clostridium botulinum 5 NCTC 7272
- Clostridium tyrobutyricum
- Lactobacillus confusus NCDO 1586
- Lactobacillus delbrueckii subsp. bulgari
- Lactococcus lactis subsp. cremoris DSM 20069
- Lactococcus lactis subsp. lactis 1 DSM 20481
- Lactococcus lactis subsp. lactis 2
- Leuconostoc carnosum
- Leuconostoc mesenteroides
- Leuconostoc oenos
- Leuconostoc paramesenteroides
- Listeria monocytogenes 1 ATCC 19115
- Listeria monocytogenes 2
- Mycoplasma flocculare
- Mycoplasma hyopneumoniae ATCC 27719
- Mycoplasma pneumoniae
- Pectinatus frisingensis DSM 20465
- Peptococcus niger DSM 20475
- Staphylococcus aureus ATCC 12600
- Staphylococcus carnosus DSM 20501
- Streptococcus oralis DSM 20066
- Streptococcus parauberis NCDO 2020
- Streptococcus thermophilus DSM 20617
- Streptococcus uberis NCDO 2038

CYANOBACTERIA
- Anacystis nidulans

FLAVOBACTERIA AND RELATIVES
- Flavobacterium odoratum ATCC 4651
- Flexibacter flexilis ATCC 23079

GREEN SULFUR
- Chlorobium limicola ATCC 8327

PLANCTOMYCES AND RELATIVES
- Pirellula marina

SPIROCHETES
- Borrelia burgdorferi 1
- Borrelia burgdorferi 2
- Borrelia burgdorferi 3
- Leptospira interrogans

RADIORESISTANT MICROCOCCI AND RELATIVES
- Thermus thermophilus

THERMOTOGALES
- Thermotoga maritima

EUKARYA

ANIMALIA

CHORDATA

VERTEBRATA

MAGNOLIOPSIDA
- Arabidopsis thaliana
- Brassica napus

Table 1. (*cont.*)

- **AMPHIBIA**
  - Xenopus borealis
  - Xenopus laevis 1
  - Xenopus laevis 2
- **MAMMALIA**
  - Homo sapiens
  - Mus musculus
  - Rattus norvegicus
- **UROCHORDATA**
  - Herdmania momus
- **ARTHROPODA**
  - **INSECTA**
    - Aedes albopictus
    - Drosophila melanogaster
- **NEMATODA**
  - **SECERNENTEA**
    - Caenorhabditis elegans

- **FUNGI**
  - **ZYGOMYCOTINA**
    - **ZYGOMYCETES**
      - Mucor racemosus
  - **ASCOMYCOTINA**
    - **HEMIASCOMYCETES**
      - Candida albicans
      - Saccharomyces cerevisiae
      - Schizosaccharomyces pombe 1
      - Schizosaccharomyces pombe 2
    - **UNCERTAIN AFFILIATION**
      - Pneumocystis carinii
  - **BASIDIOMYCOTINA**
    - **HETEROBASIDIOMYCETES**
      - Cryptococcus neoformans 1
      - Cryptococcus neoformans 2

- **PLANTAE**
  - **MAGNOLIOPHYTA**
    - **LILIOPSIDA**
      - Oryza sativa
      - Citrus limon
      - Fragaria ananassa
      - Lycopersicon esculentum
      - Sinapis alba

- **PROTOCTISTA**
  - **APICOMPLEXA**
    - **COCCIDIA**
      - Toxoplasma gondii 1
      - Toxoplasma gondii 2
      - Toxoplasma gondii 3
  - **CHYTRIDIOMYCOTA**
    - **OOMYCOTA**
      - Phytophthora megasperma
  - **CILIOPHORA**
      - Tetrahymena pyriformis
      - Tetrahymena thermophila
  - **DICTYOSTELIDA**
      - Dictyostelium discoideum
  - **DINOFLAGELLATA**
      - Prorocentrum micans
  - **EUGLENIDA**
      - Euglena gracilis
  - **PLASMODIAL SLIME MOLDS**
    - **MYXOMYCOTA**
      - Didymium iridis
      - Physarum polycephalum
  - **RHIZOPODA**
    - **LOBOSEA**
      - Entamoeba histolytica
  - **ZOOMASTIGINA**
    - **DIPLOMONADIDA**
      - Giardia ardeae
      - Giardia intestinalis
      - Giardia muris
    - **KINETOPLASTIDA**
      - Crithidia fasciculata
      - Trypanosoma brucei

PLASTIDS

- **PLANTAE**
  - **BRYOPHYTA**
    - **MARCHANTIOPSIDA**
      - Marchantia polymorpha
  - **MAGNOLIOPHYTA**
    - **LILIOPSIDA**
      - Oryza sativa
      - Zea mays
    - **MAGNOLIOPSIDA**
      - Alnus incana
      - Conopholis americana
      - Epifagus virginiana 1
      - Epifagus virginiana 2
      - Nicotiana tabacum 1
      - Nicotiana tabacum 2
      - Pisum sativum

- **PROTOCTISTA**
  - **CHLOROPHYTA**
    - **CHLOROPHYCEAE**
      - Nanochlorum eucaryotum
      - Chlamydomonas eugametos
      - Chlamydomonas frankii
      - Chlamydomonas geitleri
      - Chlamydomonas gelatinosa
      - Chlamydomonas humicola
      - Chlamydomonas indica
      - Chlamydomonas iyengarii
      - Chlamydomonas komma
      - Chlamydomonas mexicana
      - Chlamydomonas moewusii
      - Chlamydomonas pallidostigmatica
      - Chlamydomonas peterfii
      - Chlamydomonas pitschmannii
      - Chlamydomonas reinhardtii
      - Chlamydomonas sp.
      - Chlamydomonas starrii
      - Chlamydomonas zebra
      - Chlorella ellipsoidea
  - **PHAEOPHYTA**
      - Pylaiella littoralis
  - **EUGLENIDA**
      - Astasia longa
      - Euglena gracilis 1
      - Euglena gracilis 2
      - Euglena gracilis 3
      - Euglena gracilis 4
  - **RHODOPHYTA**
      - Palmaria palmata

MITOCHONDRIA

- **ANIMALIA**
  - **CHORDATA**
    - **VERTEBRATA**
  - **MOLLUSCA**
    - **BIVALVIA**

Table 1. (*cont.*)

MAMMALIA
Aepyceros melampus
Antilocapra americana
Balaenoptera musculus
Balaenoptera physalus
Bos taurus
Boselaphus tragocamelus
Capra hircus
Cephalophus maxwelli
Cervus unicolor
Damaliscus dorcas
Didelphis virginiana
Gazella thomsoni
Halichoerus grypus
Homo sapiens 1
Homo sapiens 2
Homo sapiens 3
Hydropotes inermis
Kobus ellipsiprymnus
Madoqua kirki
Muntiacus reevesi
Mus musculus
Odocoileus virginianus
Phoca vitulina
Rattus norvegicus 1
Rattus norvegicus 2
Tragelaphus imberbis
Tragulus napu

AVES
Anas platyrhynchos
Cairina moschata
Gallus gallus

AMPHIBIA
Rana catesbeiana
Xenopus laevis

OSTEICHTYES
Crossostoma lacustre
Cyprinus carpio
Neoceratodus forsteri 1
Neoceratodus forsteri 2
Protopterus sp.
Latimeria chalumnae

ECHINODERMATA
ECHINOIDEA
Paracentrotus lividus
Strongylocentrotus purpuratus

ARTHROPODA
MALACOSTRACA
Artemia franciscana
Artemia salina

INSECTA
Aedes albopictus
Apis mellifera
Apis mellifera ligustica
Drosophila melanogaster
Drosophila yakuba
Locusta migratoria
Spodoptera frugiperda
Mytilus edulis

NEMATODA
SECERNENTEA
Ascaris suum
Caenorhabditis elegans

FUNGI
ASCOMYCOTINA
HEMIASCOMYCETES
Saccharomyces cerevisiae 1
Saccharomyces cerevisiae 2
Schizosaccharomyces pombe

PLECTOMYCETES
Aspergillus nidulans
Penicillium chrysogenum

PYRENOMYCETES
Neurospora crassa
Podospora anserina

PLANTAE
BRYOPHYTA
MARCHANTIOPSIDA
Marchantia polymorpha

MAGNOLIOPHYTA
LILIOPSIDA
Triticum aestivum
Zea mays

MAGNOLIOPSIDA
Oenothera berteriana

PROTOCTISTA
APICOMPLEXA
HEMATOZOA
Plasmodium falciparum

CHLOROPHYTA
CHLOROPHYCEAE
Chlamydomonas eugametos
Chlamydomonas reinhardtii
Prototheca wickerhamii
Scenedesmus obliquus

CILIOPHORA
Paramecium aurelia
Paramecium primaurelia 1
Paramecium tetraurelia 2
Tetrahymena pyriformis 1
Tetrahymena pyriformis 2

ZOOMASTIGINA
KINETOPLASTIDA
Crithidia fasciculata
Crithidia oncopelti
Herpetomonas mariadeanei
Herpetomonas megaseliae
Herpetomonas muscarum
Herpetomonas samuelpessoai
Leishmania tarentolae
Leptomonas sp.
Trypanosoma brucei 1
Trypanosoma brucei 2

[a]In some cases, species names are listed several times followed by a sequential number, because multiple LSU rRNA sequences have been determined, usually by different authors. These sequences are not necessarily the same because they may originate from different varieties or strains, or from different genes, of the same species. The taxonomy followed for the three domains Eukarya, Archaea, and Bacteria, is as explained in the text. Plastidial and mitochondrial structures are listed according to the systematics followed for the host organism. In the case of Archaea and Bacteria, the species name is followed by the culture collection name and number if specified by the author.

differences in length of up to 900 bases. The structure of some of these regions has not yet been conclusively determined. The alignment and proposed secondary structure of the mitochondrial LSU rRNAs is less dependable because of the larger variability in both length and sequence.

The secondary structure of the molecule is treelike, with the helices forming branches which end either in a hairpin or in a multibranched loop. The stem of the tree joins the 5′ and 3′ end of bacterial LSU rRNAs. From this stem emanates a central multibranched loop. In Eukarya, and probably in Archaea the

stem helix is not present, but the central loop is. The following provisional helix numbering system is used in Figs 1 and 2. Structures branching from the central loop are labeled A to I, starting with the stem helix. Within each of these structures, helices bear a different number when they are separated by a multibranched loop. All numbering is sequentially from 5′ to 3′. Structural elements specific to certain taxa are named after the preceeding core helix followed by an underscore and number. The helix numbering may have to be revised if additional structural elements are identified in the future.

AVAILABILITY AND FORMAT OF THE DATABASE

The LSU rRNA database will be made available through anonymous ftp on the server uiam3.uia.ac.be (143.169.8.1). The files will also be made available to the EMBL nucleotide library for distribution. Researchers who cannot obtain the database through these channels, can request the database or parts thereof on magnetic media from the authors. The authors can be contacted by electronic mail to dwachter@reks.uia.ac.be or rrna@reks.uia.ac.be. On the server, a file called 'readme' will be present which describes the latest state of the database, giving the contents of the files and directories, and a description of the programs available for format conversion, alignment editing (2) and phylogenetic tree construction (16).

In order to simplify access to the database, each sequence is stored in a separate file, together with information about this sequence. The names of these files are produced from the species name by taking characters of the genus and species names. These are preceded by a code describing the phylogenetic group to which the species belongs. This makes it possible to either retrieve specific sequences using the full file name, or to retrieve a set of sequences belonging to a phylogenetic group using wild cards. Several sequence files can be integrated into one alignment using a program available on the server.

The format of the files is very simple, so that the files can be used readily by computer programs, or can easily be converted to formats used by specific programs. The files start with a few header lines which contain data about the sequence such as the accession number and literature reference. These are followed by the organism name. The sequence comes next. It consists of a range of nucleotide symbols interspersed with gap symbols necessary for alignment. The sequence end is indicated by an asterisk. The beginning and end of secondary structure elements are indicated by insertion of special symbols. Special 'helix numbering' files are present for researchers who wish to use the secondary structure information. When these are incorporated into an alignment, they indicate the name of each different helix segment.

When a sequence consists of several fragments resulting from processing, or of several exons, the sequence of each part ends with an asterisk, and has its own header containing the accession number, literature reference and a description of the sequence segment. However, the segments are stored in the same file and have the same organism name.

Users of the database are requested to cite this paper.

ACKNOWLEDGEMENTS

Peter De Rijk is research assistant of the National Fund for Scientific Research. Our research was supported by the Programme on Interuniversity Poles of Attraction (contract 23) of the Office for Science Policy Programming of the Belgian State, and by the Fund for Collective Fundamental Research.

REFERENCES

1. Rice,C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) *Nucleic Acids Res.* **21**, 2967–2971.
2. De Rijk P. and De Wachter, R. (1993) *Comput. Applic. Biosci.*, **9**, 735–740
3. Van de Peer,Y., Van den Broeck,I., De Rijk,P. and De Wachter,R. *Nucleic Acids Res.*, this issue
4. Brusca,R.C. and Brusca,G.J. (1990) Invertebrates, Sinauer Associates, Inc. Sunderland.
5. Cronquist,A. (1971) Introductory Botany, Harper & Row, New York.
6. Ainsworth,G.C., Sparrow,F.K. and Sussman,A.S. (1973), The Fungi: an Advanced Treatise, Academic Press, New York, Vol. 4A.
7. Moore,R.T. (1988) in Moriarty,Ch. (ed.), Taxonomy putting plants and animals in their place. Royal Irish Academy, Dublin, pp. 61–88.
8. Margulis,L., Corliss,J.O., Melkonian,M. and Chapman,D.J. (eds.) (1990) Handbook of Protoctista., Jones and Bartlett Publishers, Boston.
9. Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.* **21**, 3025–3049.
10. Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
11. Benson,D., Lipman,D.J. and Ostell,J. (1993) *Nucleic Acids Res.* **21**, 2963–2965.
12. Woese, C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
13. Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.* **176**, 1–6.
14. Veldman,G.M., Klootwijk,J., de Regt,V.C.H.F., Planta,R.J., Branlant,C., Krol,A. and Ebel,J.-P. (1981) *Nucleic Acids Res.*, **9**, 6935–6952
15. Michot,B., Hassouna,N. and Bachellerie,J-P. (1984) *Nucleic Acids Res.*, **12**, 4259–4279
16. Van de Peer,Y. and De Wachter,R. (1993) *Comput. Applic. Biosci.*, **9**, 177–182